Linking Spatial Named Entities to the Web of Data for Geographical Analysis of Historical Texts

Article *in* Journal of Map & Geography Libraries · January 2017

DOI: 10.1080/15420353.2017.1307306

CITATIONS
2

READS
127

3 authors, including:

Pierre-Henri Paris
Sorbonne Université
7 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Carmen Brando
École des Hautes Études en Sciences Sociales
26 PUBLICATIONS **165** CITATIONS
SEE PROFILE

Linking spatial named entities to the Web of Data for geographical analysis of historical texts*

Pierre-Henri Paris, Nathalie Abadie & Carmen Brando

Abstract: In our work, we are interested in facilitating the exploration by scholars of the geography of texts, in particular historical narrative texts describing routes. Semantic annotation constitutes the first step to enrich such text with the necessary information for producing analytical maps. The present article focus on the disambiguation of spatial named entities (SNE) by the attribution of an identifier of the ever-growing Web of Data. This giant knowledge base (KB) provides qualitative spatial information about geographic entities, in particular spatial relations such as (:Paris :southOf :Lille), (:Paris :country :France). We thus propose a graph matching algorithm relying on the A* algorithm and graph edit distances for choosing the best referent in the KB for the SNE. We perform preliminary experiments and note the clear gain in performance, we also propose some examples of maps which are built semi-automatically. Finally, we draw conclusions and describe our plans of future work.

Keywords: Web of Data, named-entity resolution, spatial named entity, historical text, geographical analysis

1 Introduction

The study of historical texts describing geographical phenomena and the projection of such evidence onto maps constitute fundamental scholarly work within Historical Studies. Novel developments of computational approaches in natural language processing (NLP) and geographic information science (GIS) aim to facilitate the

*https://www.tandfonline.com/doi/full/10.1080/15420353.2017.1307306

geographical analysis of historical phenomena by scholars, specially when dealing with significant amounts of texts. For instance, population mortality and disease spread in nineteenth century Great Britain (Porter et al. 2015; Porter et al. 2016) as well as the journeys of commercial travellers in sixteenth century Europe (Blank and Henrich 2016) are analyzed through these methods, as similarly done with distant reading approaches in literary texts (Moretti 2007).

The analysis of geographical phenomena begins with the spotting of mentions to historical places in our texts. Any place mention refers to spatial named entities (SNE) which are automatically recognized through named entity recognition (or geoparsing) systems. The detection is followed by a classification phase according to a place taxonomy (see for instance the categories proposed by the Ester 2 annotation guidelines[1]). Afterwards, named-entity resolution (also known as named-entity linking and henceforth referred to as NEL) associates a named entity tagged in text with the identifier of its referent belonging to a Knowledge Base (KB). KB are usually exposed as Linked Open Data (LOD)[2], and are interconnected to further LD sets by equivalence links (for instance, owl:sameAs and skos:exactMatch). In this context, the aforementioned identifiers are known as IRI[3]. Besides, linking SNE to a LD set provides opportunities to enrich a text with external information which can be automatically retrieved by dereferencing the IRIs or querying the LOD sources. Many customizable applications can be built on top of these new data for querying and visualizing information related to the text (Frontini, Brando, and Ganascia 2016).

[1] (in French) http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

[2] http://lod-cloud.net/

[3] Acronym which stands for Internationalized Resource Identifier

The present article focuses on NEL and presents a graph- and knowledge- based approach which relies on LD and most importantly on spatial relations, in particular directional, topological or metric relations, to disambiguate SNE. These kinds of relations seem considerably useful for our purposes as geographical locations are often unknown when dealing with historical places. They describe instead relative positioning among places and the overall spatial configuration of the geographical space (Frank 1996; Cohn et al. 2014). Hence, reasoning on such spatial graph helps us retrieve most accurately locations of historical places. In this article, we work on the historical essay “La grande peur de 1789” (Lefebvre 1932) written by French historian Georges Lefebvre (1874-1959) which describes an astonishing event which took place in France between July 20, 1789 and August 6, 1789 in the rise of the French Revolution. A collective feeling of fear spread throughout the French countryside at the idea that aristocrats would ravage the crops and harm the people. In particular, the author describes the spreading of the fear in narrative manner and details the way fear travels from Paris to the entire French territory. In other words, the fear follows somehow several routes with different start points, it expands diachronically as a network while crossing several paths until arriving to multiple places. In addition, our text is the object of a digital edition in XML-TEI, so as to be compliant with the recommendation of the Text Encoding Initiative[4], this is essential for any digital humanities project.

Finally, the remainder of the article is organized as follows. Section 2 describes related work on resolution of spatial named entities. Section 3 presents our graph matching algorithm as well as the ontological and the geographical KB in LD we produced. Section 4 describes our experimentation and preliminary results. Section 5 presents different maps we created focusing on the old French province of Champagne

[4] http://http://www.tei-c.org/

and the South West of France which are the object of our study. Finally, we draw conclusions and state our ideas of future work in section 6.

2 Related work

According to (Leidner 2007) and (Smith and Mann 2003) there are several types of ambiguities. The one we chose to spot in our work is the so-called referent ambiguity. A referent ambiguity occurs when a place name is shared by two or more locations (e.g. the name “Paris” can refers to more than twenty locations across the world). We spotted three sorts of approaches for SNE resolution, see (Buscaldi 2011) for a detailed survey.

Firstly, coordinate based approaches which are inspired by Tobler's first law of geography: “everything is related to everything else, but near things are more related than distant things.”. These approaches rely on unambiguous and georeferenced SNE that are present in the text to disambiguate ambiguous ones. (Smith and Crane 2001; Buscaldi and Rosso 2008) calculate the centroid obtained from coordinates of both the candidates of the SNE and the unambiguous SNE, to choose the best candidate. In (Derungs and Purves 2014; Zhao et al. 2014), from the coordinates of the unambiguous SNE, a score is given based on the distance between the latter and the ambiguous SNE candidates. When an unambiguous SNE *N* is chosen, for each candidate l_i of an ambiguous SNE *T*, a score l_i is given as a function of the distance between *S* and l_i, which makes it possible to order the candidates and to select the one that has the best score as referent of *T* . Other approaches used the clustering algorithm DBSCAN to create clusters based on spatial density (Intagorn and Lerman 2011; Moncla et al. 2014). This family of approaches is historically the first to have been explored since it is intuitive.

Secondly, the data-driven approaches relying on machine learning algorithms. According to (Buscaldi 2011), they use the non-geographical content to construct probabilistic models by using relations between entities (e.g. persons, organization, etc.) and places. For example, if the word “Google” is found in “Mountain View”’s context, it may suggest that the right reference place is Mountain View, CA rather than Mountain View, Arkansas. Many machine learning algorithms have been used by these approaches: Naive Bayes classifier (Smith and Mann 2003), Gradient Boosted Decision Tree (Agrawal and Shanahan 2010), SVM (Martin et al. 2010), Conditional Random Fields (Habib and Van Keulen 2012), or LambdaMART (Santos et al. 2015). Results are often better than those of coordinate-based approaches, but data-driven approaches main drawback is the need of a big enough and context relevant annotated corpus. It may be very difficult to obtain one for a specific domain.

Thirdly, knowledge based approaches depend on knowledge sources like Wikipedia, DBpedia or Yago to determine if SNEs related (in the KB) to the candidate of an ambiguous SNE are present in the document. The KB as well as the document are used to seek additional information about the ambiguous SNE (like size, population, document’s creator, etc.) in order to facilitate the disambiguation process with the help of external knowledge. (Amitay et al. 2004) propose an algorithm (named Web-a-Where) that uses population statistics and hierarchical relationship between SNE from the KB (a gazetteer built upon multiple data sources). For example, if “Paris” and “Dallas” are quoted in a text, those two toponyms having the state of Texas as common ancestor in the KB, “Paris” is most likely to be about Paris, Tx than Paris the French capital. (Overell and Rüger 2008) propose an algorithm using additional information such as size, importance or inhabitants number of a place to rank candidates (for example a location with more inhabitants is more likely to be quoted in a text). The knowledge source used is Wikipedia. (Buscaldi and Rosso 2008) use WordNet to compute the conceptual

density (correlation measure between the sense of a given word and its context) for each given ambiguate SNE candidate. The chosen candidate is the one that maximize the density measure. To note that the smaller the context (i.e. sentence instead of full document), the better the result is. The algorithm of (Ireson and Ciravegna 2010) is designed for documents extracted from social networks. It uses external knowledge such as users comments to find clues for helping to disambiguate. Yahoo! GeoPlanet is the KB used to measure the semantic proximity between SNE. (Bensalem and Kholladi (010) use WordNet to measure the hierarchical proximity between toponyms. (Batista et al. 2012) propose an algorithm that uses two approaches based on semantic similarity between two ambiguate SNE. For example, if a press article quotes “Lisbon” and “Porto”, as they can be both city, Lisbon is more likely to be a city than a street. The KB used is Geo-Net-PT. (Speriosu and Baldridge 2013) algorithm idea is to extend the context (local or global) by using non geographical terms. For example, if a text contains the terms “lobster”, then the toponym “Portland” is more likely to be Portland, Maine than Portland in Michigan or Oregon. The KB used is Wikipedia. (Daiber 2013) propose a nine languages supported tool called Dbpedia-Spotlight. The algorithm uses a probabilistic model inspired by (Han and Sun 2011). The idea is to rank formerly selected candidates by using a Term Frequency-Inverse Document Frequency variant. The first candidate should be the ambiguate SNE referent.

Furthermore, a variation of knowledge-based approaches are graph-based ones. The latter are relatively new and also rely on a KB exclusively structured as LD (i.e. graph data in RDF). A graph is built from the candidates selected in the KB for each mention and graph analysis operations are applied to it to determine the referent of each of the mentions. In general, when several mentions are present in the context (sentence, paragraph, whole document, etc.), the resolution of the mentions uses the other mentions that are present. These approaches exploit knowledge found in KB in order to

rank candidates by using relevance hypothesis (centrality, semantic similarity, additional data from the text, etc.). There are several tools that use a graph-based approach for NE resolution, such as AIDA by (Yosef et al. 2011), NERSO by (Hakimov et al. 2012) or REDEN by (Brando, Frontini, and Ganascia, 2015). The general principle is to use the notion of node importance to identify which candidate will most likely be the referent of a mention in the graph of candidates. Several methods can be used to determine the importance of a node, such as the centrality score (Usbeck 2014), semantic proximity (Moro et al. 2014), a voting system (Ferragina and Scaiella 2010), etc. As (Brando, Frontini, and Ganascia 2016) point out, most of these approaches are entirely dependent on Wikipedia or DBpedia as knowledge bases. Remarkably, Babelfy is an algorithm proposed by (Moro et al. 2014). It uses a graph to disambiguate SNE with a heuristic help that founds subgraph of maximal density in the BabelNet KB (nodes are toponym candidates) and selects coherent semantic interpretations to find the referents. (Brando, Frontini, and Ganscia 2016) propose an algorithm using the notion of centrality in Graph Theory. Many measures exist to measure a node centrality. For example, the degree centrality (number of neighbors possessed by a node), betweenness centrality (number of times a node acts as a bridge along the shortest path between two other nodes), or Eigenvector centrality (influence of a node), etc. The idea is to build a graph with the candidates and to calculate its nodes centrality to keep the most salient node as referent.

Finally, our approach can be considered as both a graph- and knowledge- based approach. We develop more precisely a graph matching algorithm which disambiguates SNE relying on the knowledge of some LOD cloud sources and more importantly on explicit spatial relations. This is the subject of the next section.

3 Qualitative spatial graph matching for the resolution of SNE

Our approach focuses on the SNE resolution task. The recognition (i.e. detection and classification) and other tagging tasks are previously performed by external natural language processing (NLP) tools. Indeed, the approach is meant to be integrated in a chain of specialized NLP tools within any digital humanities project interested in textual corpora semantic annotation. This is also the reason why it is TEI-compliant (essentially the “placeName" tag[5]).

Our resolution approach relies on spatial relations, in particular directional, topological or metric relations. Indeed, we cannot only depend on names of historical places because they may have changed through time. In general terms, we disambiguate the SNE tagged in the text by comparing their labels, possible nature, and spatial relations evoked in the text with the candidates present in the KB. To do so, the text needs also to be annotated in cardinal orientations, verbs of movement and dates, whether manually or semi-automatically (including manual correction) by a specialized semantic tagging tool such as Perdido (Moncla 2015)[6] (see example 2).

Furthermore, if two place names resemble each other and have the same spatial relations with the same other place names, then the chances for them to designate the same place are strong. So to disambiguate candidate toponyms, we propose to compare their spatial relations in the KB with those found in the texts to be treated and which describe the relative positions of the SNE mentions in these texts. We thus develop a graph matching algorithm which compares the graph of spatial relations existing among SNE in the text and all the graphes of spatial relations existing among candidates in the

[5] http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-placeName.html
[6] http://erig.univ-pau.fr/PERDIDO/

KB. The goal is to find the subgraph within the KB containing the best candidates i.e. candidates having the same spatial configuration as SNE described in the text.

Moreover, our approach leans on the best practices of Linked Data and the Semantic Web promoted by the W3C Consortium. It is KB independent, in other words, any KB in LD which describes geographic features, their nature (e.g. lake, city), and the spatial relations between them (e.g. north of, within) can be plugged to our algorithm.

We illustrate the idea behind our approach with an excerpt (in French) below of the text “La grande peur de 1789” (Lefebvre 1932) tagged in SNE (in bold) and in other meaningful annotations such as nature, orientations and so on (in italic). Our aim is to disambiguate the tagged SNE by attributing to each entity the proper identifier (an IRI) of the KB.

Example 1:

“Mais c'est *vers le Sud* que le mouvement *parti de* **Ruffec** remporta le plus de succès en submergeant presque toute *l'***Aquitaine**. Il *descendit* d'abord *la* **Charente** *par* **Mansle** et fut *à* **Angoulême** dès le 28 à trois heures du soir ; il *suivit le fleuve par* **Jarnac** *et* **Cognac** jusqu'*à* **Saintes** ; là, on perd sa trace et *la* **Saintonge** maritime et méridionale semble être demeurée tranquille.”

“Mais c'est *vers le Sud* que le mouvement *parti de* **Ruffec** remporta le plus de succès en submergeant presque toute *l'***Aquitaine**. Il *descendit* d'abord *la* **Charente** *par* **Mansle** et fut *à* **Angoulême** dès le 28 à trois heures du soir ; il *suivit le fleuve par* **Jarnac** *et* **Cognac** jusqu'*à* **Saintes** ; là, on perd sa trace et *la* **Saintonge** maritime et méridionale semble être demeurée tranquille.”

In example 1, all tagged SNE are ambiguous. Indeed, there exist at least two and up to ten candidates in DBpedia for each one only by retrieving from the KB the places having similar names. In this context, one may argue that metric proximity would alone

help to determine the best referents for the SNEs in our text. In fact, neither coordinates nor metric distances are reliable indicators to answer such task. On the one hand, coordinates are often unknown in the KB typically when dealing with historical places . Moreover, geographic feature with a large spatial extent or a linear shape are still often represented by caracteristic points in LOD sources: for example, watercourse are described in DBpedia by points locating their source or mouth. Disambiguation approaches based on spatial proximity require candidates whose locations are described precisely enough in the KB so that they can reflect the local geographic context described in the text. As such precise geometries are not always available in LOD sources, it is hence necessary to capture more evidence, such as relative cardinal direction between two places, in order to make the best decision.

In this sense, we notice the presence of the cardinal relation “sud” (south) as well as other ones that can be extracted using lexico-syntactic patterns. In our example, as humans, we can clearly detect points of departure (e.g. “Ruffec”) and points of crossing and arrival (e.g. “Mansle”, “Angoulême”, “Jarnac”). There are thus four possible spatial relations, i.e. “Mansle” south of “Ruffec”, “Angoulême” south of “Ruffec”, and so on. We aim at trying to automatize detection of such relations using heuristics. Also, the nature of the geographic entity may provide interesting insights to the disambiguation process. Indeed, a river “A” which crosses a city “B” can be translated into a proximity relation between both entities. Moreover, the early identification of the nature of the SNE also helps to reduce the number of candidates and thus the overall ambiguity. The semantics of some words from the context around an SNE may sometime help to deduce such knowledge. It is worth mentioning that we haven’t conducted any experiments giving us statistics about our attempts to find those relationships. Our heuristics are patterns, developed in XSLT, using a combination of small lexicon (“valley”, “river”, “mont”, etc.), annotations from TEI and simple syntactic rules. For example, if an SNE is preceded by

“la” (“the”) and we found the word “river” some words before, we assume the SNE may belongs to the type “bodyOfWater”. Because of the verbosity of the XSLT language we choose to not include any example.

Example 2 :

```
    <w lemma= »vers » type= »PREP »>vers</w>
    <term subtype= »orientation » type= »offset »>
      <w lemma= »le » type= »DET »>le</w>
      <w type= »NPr »>Sud</w>
    </term>
    <w lemma= »que » type= »CONJC »>que</w>
    <w lemma= »le » type= »DET »>le</w>
    <w lemma= »mouvement » type= »N »>mouvement</w>
    <phr type= »motion »>
                    <w   lemma= »partir »   subtype= »motion_initial »
type= »V »>parti</w>
      <placeName>
                      <certainty   assertedValue= »rs »   degree= »0.5 »
locus= »name » />
        <w lemma= »de » type= »PREP »>de</w>
        <name xml :id= »146 »>
          <w lemma= »Ruffec » type= »NPr »>Ruffec</w>
        </name>
      </placeName>
    </phr>
```

Example 2 shows a part of example 1 in TEI format. We can observe that “sud” (“south”) is tagged as an orientation, “parti” (“leave”) is tagged as an initial motion verb. So there must be a spatial relationship of the type <:? :southOf :Ruffec>.

Having the aforementioned ideas in mind, we developed the approach which is summarized in Figure 1. The input parameter is the tagged TEI file. The RDF graph is built by extracting the SNE and spatial relations from the text and describing them consistently with two ontologies designed for that purpose (see subsection 3.1). At the same time, it is necessary to set up the KB (see subsection 3.2). Next, the algorithm

analyses the structure of graph and produces several matches between subgraphs coming from the graph built from the text and the graph of the KB (see subsection 3.3). Finally, these mappings are used to disambiguate the SNEs contained in the input TEI/XML file and the right referents are written out to the same file (see subsection 4.5). Finally, subsection 3.5 describes an module to transform the enriched XML/TEI file into routes serialized in RDF/XML. These structured data are interpreted by a cartographic application that we developed to visualise the locations of the resolved SNE and their properties.

[Insert Figure 1 here: Our approach]

3.1 Building RDF sequences of routes from TEI annotations

Above all, our work is built upon four languages recommended by the W3C, namely RDF, RDFS, OWL and SPARQL. RDF are graph triplets of the form "subject", "predicate", "object". The subject is a resource corresponding to everything that can be referred to: a person, a place, a web page, an image, etc. The predicate is a property that describes that resource, and the object is the value of that property for that resource. RDF Schema (RDFS) and Web Ontology Language (OWL) are knowledge representation languages built on the RDF data model and based on the description logic formalism. They enable the design of ontologies, in other words, a structured set of terms and concepts representing the meaning of an information field. Most importantly, modeling of knowledge makes hence possible to infer new facts from known facts. Those languages are the fundamental bricks for the KB in LD.

The input TEI file is transformed into an RDF graph which contains as much spatial relationships among the SNE identified in the text as possible (as we use heuristics, some relationships may escape us). We defined several lexico-syntactic rules

to deduce spatial relations between SNE (A is south of B, C within D). The text describing the routes, the nature and dates associated with the SNE in text are also taken into consideration. For example, the following text (SNE in bold and other annotations in italic) “Vers l’*Ouest* de **Ruffec**, à **La Rochelle**” (to the *west* of **Ruffec**, to **La Rochelle**) becomes the RDF triplet <:La_Rochelle :westOf :Ruffec> and the text “*la rive* gauche de la **Seine**” (the left *bank* of the **Seine**) becomes the RDF triple <:Seine rdf:type :Rivière>. Furthermore, the spatial relations as well as the structure of the routes described in the text are defined by two ontologies created in order to describe the spatial relations between SNE and to formalise the structure of the routes extracted from the text.

For the ontology of spatial relations, we first sought out the different potential candidate ontologies listed in the catalog of Linked Open Vocabularies (LOV)[7]. The ontology TISC[8] contains cardinal spatial relations of interest (ex: "tisc:southEastOf".) but several useful axioms are missing. We chose thus to extend it by adding the missing owl:inverseOf relationship between TISC properties of interest for our application (for example "north" is the inverse of "south" and vice versa) . We also added the RDFS construct rdfs:subPropertyOf to the relations composed of two cardinal directions. This enable for us to have the possibility to either use “A northOf B” or “B southOf A” during the matching step. For example ”rlsp:northEastOf” is therefore defined as an extension of "tisc:northEastOf" and is the inverse relation of "rlsp:southWestOf" (to be able to navigate in the desired direction subsequently when matching the graphs and the shortest paths).

For instance, the relation ”north-est of “ (compose of “north” and “est”) is defined in the following manner (serialized in RDF/XML):

[7] http://lov.okfn.org/dataset/lov/

[8] http://observedchange.com/tisc/ns/

rlsp:northEastOf a owl:ObjectProperty;

rdfs:label "au nord-est de"@fr, "north-east of"@en;

rdfs:comment "Property defining the relative position of a geographic entity located at the north-est of another geographic entity."@en;

rdfs:isDefinedBy rlsp:northEastOf;

rdfs:subPropertyOf tisc:northEastOf;

rdfs:subPropertyOf rlsp:northOf;

rdfs:subPropertyOf rlsp:eastOf;

owl:inverseOf rlsp:southWestOf.

Likewise, we have defined an inclusion property (which derives from the ontology OSSPR) to signify that a resource representing a place contains another resource representing another place (e.g. a city in a region).

The ontology of routes defines the structure of the RDF graph obtained by means of some XSLT transformations applied to the text in the TEI/XML format. It describes the concepts and properties necessary for the representation of routes or phenomena propagating in space. The RDF graph is composed of RDF sequences (rdf:Seq). Sequences contain routes. A route is an ordered list of waypoints. A waypoint can have several properties such as the date it is reached, the date it is left, or a spatial reference that ideally makes it possible to locate it unambiguously. The “spatial reference” property associates firstly to each waypoint its SNE mention extracted from the TEI/XML file after the XSLT transformations.

The RDF graph thus created is composed of several sequences (rdf:Seq). A sequence is defined as a sequence of SNEs concerned by the same cardinal orientation. Indeed, a phenomenon may have an epicenter and thus propagate concomitantly in several different directions. For example, if the phenomenon starts from

A and propagates westward to B then C and at the same time from A to east to D then we will have two sequences (A, B, C) and (A, D) (see Figure 2).

[Insert Figure 2 here: Example of sequence creation]

Each sequence contains routes that are composed of the SNEs within the same sentence. We therefore segment the text into sentences relying on punctuation as it is usually performed in written text. We assume that SNE mentioned in the same sentence are part of the same route. Finally, these routes are composed of bags (in the sense of rdf[9]:Bag) which contain the SNE joined by conjunctions or by a comma. Let us take the following example : "... north of A to B and C before reaching D.", it is impossible to know whether the phenomenon has reached B and C simultaneously or not and if it is left from B and C to D or from C to D (see Figure 3). In other words, a bag can be understood as a step within a route, a set of waypoints.

[Insert Figure 3 here: Principle of bags of SNE]

We consider each of these sequences as a context in their own right. We assume that two different SNEs with the same name can occur in the same text even if it goes against intuition. The reason is we want to keep our approach as generalist as possible. It would be very interesting to assume the opposite in future work. The homonymy is one of the most frequent and complex types of ambiguity to be solved according to Leidner (2007). The sequences are processed one by one and considered as an independent RDF subgraph which is then matched to the KB graph to proceed with the resolution of the SNE. Figure 4 shows the sequence from the excerpt presented in section 3.

[9] https://www.w3.org/TR/rdf-schema/

[Insert Figure 4 here: Complete example of sequence]

It should be noted that if the graph of the input sequence of the matching algorithm contains alternatives (rdf:Alt), that is, if for one (or more) routing point(s) the heuristics have proposed natures of different geographical entities (eg "mountain" or "river"), then we will create a graph by possibility, and each of these graphs will be paired with that of the knowledge base (see Figure 5).

[Insert Figure 5 here: Two graphs to match due to the presence of an alternative in the sequence]

We implemented transformations from TEI file to RDF graph by using the eXtensible Stylesheet Language Transformations (XSLT). The purpose of this functional language is to transform an XML document into another XML document or text document. The transformation takes place in several distinct stages in order to separate the different objectives of each one and to facilitate the creation and reading of the proposed code. All XSLT files are available on our github repository[10].

3.2 Preparation of KB

As we have mentioned, KB are built upon the RDF graph model which is the standard for knowledge representation. It is feasible to plug any RDF-based generalistic or domain-specific KB, as proposed by (Brando, Frontini, and Ganascia 2016), according to the textual corpora at stake. We have decided to use the French version of DBpedia[11] as KB. To extract the useful knowledge we have executed a SPARQL query schematized in Figure 6. As we rely strongly on spatial relations, several improvements have been made to the KB in order to add cardinal spatial relations between the

[10] https://github.com/cvbrandoe/REDEN/tree/geo-reden
[11] http://fr.dbpedia.org/

resources typed as places and, on the other hand, topological relations between resources types as municipalities and resources typed as watercourses.

[Insert Figure 6 here: KB enrichment process]

To determine the missing cardinal spatial relations, we therefore built a Voronoi diagram based on . Then the coordinates of each "dbo:Place[12]" resource located inside the bounding box encompassing metropolitan France. For each of these resources, the coordinates are transformed into Lambert 93 projected coordinate reference system so that they can be processed consistently with the Geotools[13] algorithm that we used to calculate the Voronoi diagram. Once the diagram is computed, each voronoi cell's neighbors are retrieved. Then, for each cell-neighbor pair, we get the coordinates of the corresponding resources and we compute the azimuth between the points that locate the two resources involved. The type of cardinal relation that must be assigned to each cell-neighbor's corresponding resources pair is then determined depending on the value of their azimuth.

The intersection of watercourses with municipalities area has been calculated using vector- hydrography[14] and administrative units[15] geographic databases. This relationship has been added to the KB in the form of the triple ":river <http://en.dbpediaorg/property /villes> :city".

[12] http://dbpedia.org/ontology/Place
[13] http://www.geotools.org/
[14] http://www.sandre.eaufrance.fr/atlas/srv/fre/catalog.search#/metadata/49c7c071-7afb-4ff3-a00b-81af7425045f
[15] http://professionnels.ign.fr/geofla

3.3. Selection of candidates

This step aims to find the candidates for each SNE in the text. To do so, we use a measure of similarity between strings of characters. It is a question of comparing the name of the label on one side, and on the other, the rdfs:label or the foaf:name of the various resources in the KB in order to identify whose labels are most similar to the SNE. In fact, if only one of these two properties is present, we perform the calculation with it, if both are present, we perform the calculation with both and take the best result. In most cases, the values of these two properties are the same, but sometimes the nature of the SNE is attached to it, for example "http://en.dbpedia. org/resource/Yonne_(d %C3%A9partement)" has for rdfs:label "Yonne (department)" and for foaf:name "Yonne". Only the candidates above a certain threshold defined by the system's user, are retained, and the best *N*s are selected, *N* being also chosen by the user.

We tested several well-known measures of similarity between strings such as: cosine similarity, Jaccard index, Jaro-Winkler distance, Longest Common Subsequence distance (Bakkelund 2009), normalized Levenshtein distance, Sorensen-Dice index, Token-Wise string similarity (Khrouf et Troncy 2011)[16]. We then compared them to select the one that best suits our needs by calculating a precision score that corresponds to the percentage of candidate set containing the correct resource compared to a manually annotated reference TEI file (gold standard). All the TEI files we used have been created with the tool Perdido. Our experimentation showed that the cosine similarity always come first and hence we have chosen it (see Table 1).

Number of candidates	200	100	50	25	10
Cosinus	0,85	0,85	0,84	0,83	0,78

16

We did not reimplement these similarity measures as they are available here: https://github.com/tdebatty/java-string-similarity

Jaccard	0,85	0,84	0,84	0,79	0,75
Jaro-Winkler	0,84	0,84	0,82	0,79	0,72
LCS	0,71	0,70	0,70	0,69	0,68
Levenshtein normalisée	0,73	0,72	0,70	0,69	0,68
Sorensen-Dice	0,85	0,84	0,84	0,79	0,75
Token-Wise	0,72	0,70	0,70	0,70	0,68

Table 1: Precision scores of similarity measures
(in green : best measures for the corresponding column)

3.4 RDF graphs matching

Once the graph of spatial relations is built, we perform the matching between the aforementioned graph (i.e. composed of sequences) and the graph of the KB in order to disambiguate the tagged SNE. We chose an algorithm within the family of approaches based on graph edit distance (GED). This kind of algorithms use the notion of editing path which is a bijection for passing from one graph to another. Several edit operations (substitution, insertion and deletion) can be applied and each one has a cost. The GED is the cost of the path that has the minimum cost. More precisely, we use the algorithm of A* Beamsearch proposed by (Neuhaus et al. 2006). It is based on the algorithm A* and aims to find the GED between two graphs by gradually constructing the possible edit paths and continuing with the best (least expensive) until obtaining a complete edit path.

The advantage of this family of GED approaches is the polynomial complexity and the fact that they allow for easy optimization in our context. In other words, we can keep only a constant number of nodes to be explored, thus reducing the memory required. It allows to control to some extent the RAM to be used (as we will see this approach is greedy in terms of RAM), it is simple to implement and above all the edit path provides immediately the referents (they correspond to the substitutions). It is a question of finding the least expensive edit distance between the two graphs by

constructing it using the algorithm A* in which the number of paths is limited in order to limit the necessary memory space. The solution found is suboptimal, that is to say that there may be cheaper paths.

In opposition, the alternative two families of approaches are i) graph kernels (Gärtner et al. 2002) based on support vector machines, and ii) graph embeddings build upon a vector representation (Riesen and Bunke 2010). The former is very expensive (O (n4)) and the latter is mainly useful when performing post-processing analyzes on the embedded space. See (Livi and Rizzi 2013) for an extensive review of the three aforementioned families of graph matching algorithms.

[Insert Figure 7 here: Flowchart of our graph matching algorithm]

Our graph matching approach, based on the graph edit distance (GED) and the A* Beamsearch algorithm is presented in figure 6.1. It requires a cost function for each of the three edit operations (substitution, deletion and insertion). Let $g_{seq}=(V_{seq}, E_{seq}, \mu_{seq}, \upsilon_{seq})$ be a graph generated from a TEI sequence (see the example in Figure 4). V_{seq} is the set of nodes of g_{seq} , E_{seq} the set of arcs of g_{seq} , μ_{seq} is the function that associates an element of V_{seq} with its label, and υ_{seq} the function that associates an element of E_{seq} with its label. We denote by $t_1,\ldots,t_n$, the *n* elements of V_{seq} . Each t_i of V_{seq} has $\max_{candidat}$ candidates. These candidates being the best candidates of the knowledge base (according to a similarity score of the labels), we define the utility function which gives the shortest path between two nodes (passing only by the spatial properties "propfr: north", "propfr: south", etc.): $sp: V_{KB} \times V_{KB} \rightarrow \aleph^{+}$

The following cost functions are defined:

Deletion node cost $del : V_{seq} \rightarrow 1$

Insertion node cost $ins : V_{kb} \rightarrow 1$

The insertion and deletion costs are often determined using a machine learning algorithm. We have chosen to assign them empirically the constant value 1 as a first approximation for the time being.

Substitution cost of a node *n* of the TEI graph by a node m of the KB graph. This cost must be as low as possible when two nodes match well:

$$: V_{seq} \times V_{kb} \rightarrow [0,1] sothat \quad (n,m) = a.(1 - scoreLabel(n,m)) + b.rlsp(n,m)$$
$$+ c.link(n,m) + d.type(n,m)$$

where

- $a+b+c+d = 1 \ and \ (a,b,c,d) \in \Re^4$

- $scoreLabel : V_{seq} \times V_{kb} \rightarrow [0,1]$ gives the similarity score between the labels of the two resources. This score is precalculated when the candidates are selected (see Figure 8).

[Insert Figure 8 here: Flowchart of LABEL sub-function]

- The SPRL (for SPatial ReLation) sub-function corresponds to the score of dissimilarity between the spatial relations configuration of a node of the sequence

and the spatial relations configuration of a node of the KB. The lower this score, the more likely the node of the KB is to replace the sequence node at a low cost (see figure 6.3). The idea is that if we have two places in the text (e.g. A and B) and orientation between them (e.g. A south of B) then a candidate of A is more likely to be is referent from the KB if this candidate is at the south of the referent (or a candidate) of B. Moreover, the shorter the path between the candidates in the KB the better.

[Insert Figure 9 here: Flowchart of SPRL sub-function]

- The sub-function LINK corresponds to the dissimilarity score computed for the path between waypoints described in the text and each shortest path between candidates that can be found in the KB (see figure 6.4). As for the SPRL sub-function we seek for shortest path in the KB. We have to use this function in the substitution cost because not all SNEs are linked by cardinal or topological

relations either it is not specified in the text, either we weren't able to determine their cardinal or topological relations at the routes building step.

[Insert Figure 10 here: Flowchart of LINK sub-function]

- The sub-function TYPE : $type : V_{seq} \times V_{kb} \rightarrow [0,1]$ so that $type(n,m)=0$ if the types of *n* and *m* are identical, 1 otherwise (see figure 6.5). The idea is that a candidate should be a better one if it has the same type as the one from the SNE deduced by means of our heuristics.

[Insert Figure 11 here: Flowchart of TYPE sub-function]

$\lambda_{n,n'}$ $\lambda_{n,n'}$

Once all the sequences have been matched, the substitutions of a node A of a sequence by a node B of the KB make it possible to mark the resource represented by the node B as being the referent of the SNE mention associated with node A. Deleting a

node from the sequence means that the referent of the mention has not been found: it is therefore marked "NIL". In output, we thus produce both the original TEI file including the attributes containing the referent for each SNE, and the RDF file containing the RDF graph of the routes also enriched by the URIs of the referents identified by our approach.

That is, the number of set of candidates containing the correct resource in relation to the gold standard divided by the number of candidate sets.

4 Evaluation and results

If DBpedia does not contain the referent, then the value "NIL" is filled in instead of a referent. We measure in terms of several indicators proposed by Hachey et al. (2013) and refined by (Brando, Frontini, and Ganascia 2016). They are candidate precision, candidate recall, NIL precision, NIL recall, disambiguation accuracy and overall linking accuracy. The gold file corresponds to the excerpt (of 1537 tokens) of La Grande Peur dealing with the old French province of Champagne. Our approach has been compared to the named-entity resolution tool REDEN (see Brando, Frontini, and

Ganascia 2015), and it was executed on the gold standard and under the same conditions.

The algorithm is based on several parameters :

The RDF/XML file containing the routes

The maximum number of candidates selected for each mention of a place name

The weights used to weight the sub-functions used in the substitution cost

[Insert Figure 12 here: Number of candidates selected (with parameters label=0.3 / sprl=0.4 / link=0.1 / type=0.2)]

We ran our approach by changing the number of candidates selected for each mentions (see figure 6.5). We can see that the recall of candidates increases with the number of selected candidates but the overall linking accuracy reached its peak for 15 candidates.

Figure 13 shows the results.

[Insert Figure 13 here: Results of the comparison with REDEN]

We can observe that the precision of the candidates, that is to say the proportion of non empty candidate sets containing the correct URI with respect to the number of non empty candidate sets returned by the selection process, is better for REDEN. Only mentions with candidates are considered. REDEN has only 89 such mentions, against 179 in our case. This tends to show that REDEN has more strict selection criteria that our approach, even if this makes it lose potential good candidates. No surprisingly, REDEN has a lower score of candidates recall than our approach. The candidate recall is defined as the proportion of non-empty candidate sets containing the correct URI returned by the selection process with respect to the number of all mentions that actually have a good reference in the KB. This result shows that our approach has less strict selection criteria and therefore accepts more candidates (including good ones) during the selection step. Similarly, the precision of the "NIL" values, i.e. the proportion of empty candidate sets returned by the selection process when it is actually impossible to find a good reference in the KB, with respect to all empty candidate sets returned, is to the advantage of our algorithm because REDEN has not found a candidate for a large number of mentions while there actually was a referent in the KB (115 mentions without candidate for REDEN against 17 for us). The recall of the "NIL" values is maximum for REDEN. It evaluates the proportion of empty candidate sets returned by the selection step when the KB does not actually contain the good reference for the tagged mention, with respect to all NIL references which should actually be retrieved. Our approach is therefore more likely to assign a reference to a reference to SNE that should not have one if the disambiguation step fails. This may be avoided by improving the use of deletion and/or insertion cost functions. Finally, the overall accuracy (referent and NIL values) is 59% for our algorithm versus 50% for REDEN. To summarize, these preliminary experiments showed a clear gain in performance using our approach with the kinds of text that interest us for our study.

[Insert Figure 14 here: Comparison with different parameters (label/sprl/link/type)]

In order to choose the weight parameters used in the preceding experiments we ran our approach by changing them and finally selecting 0.4 for the SPRL function, 0.3 for the LABEL function, 0.1 for the LINK function and 0.2 for the TYPE function.

5 The geography of La Grande Peur

We developed a web- mapping application to visualize the results of the RDF graph containing the routes on top of a base map. In order to better situate these routes in their historical context, we chose to use the Cassini map scans available in a WMS service via the IGN Geoportal[17] as a base map. The green dots correspond to the waypoints whose coordinates were found in DBpedia, any orange dots correspond to waypoints whose coordinates were not available in DBpedia but were interpolated relative to neighboring waypoints: we roughly use the mean coordinates of the preceding and following bags to compute the missing coordinates. The green lines connect the waypoints in their order of appearance in a route. Each waypoint of a bag is linked to each waypoint of the preceding and following bags. Figure 15 shows an example of a route whose waypoints have been extracted from the following text, disambiguated and plotted on the Cassini base map: “Mais on la retrouve sur l'Adour, bien au Sud d'Aire, à Maubourguet et Vic-de-Bigorre où elle a dû venir de Mirande. Sur le Gers, elle est à Auch le 3 août, transmise sans doute par Lectoure.”

[Insert Figure 15 here: Example of a South West route on the web interface]

[17] https://www.geoportail.gouv.fr/

We can see on this excerpt of a sequence of routes of La Grande Peur of the South West that two of the SNEs have been poorly resolved (the two easternmost green points), as the routes described in the text are most often located in areas of a few tens of kilometers long. This may be the result of incorrect deduction from our heuristics (e.g. wrong direction in SPRL) or of bad parameters in our disambiguation algorithm. By zooming (see Figure 16) on the waypoints located in the southwestern part of the map, we can better see the two routes followed by La Grande Peur (the three points aligned along a north-south axis of one side and the others on the other).

[Insert Figure 16 here: Zoom on the southwest zone of the map of Figure 15]

Figure 17 presents another type of visualization using dates we were able to link to waypoints. The file used correspond to the output of our approach for the excerpt about the old province of the Champagne. The black dots correspond to places for which there was no date of passage in the text. The light orange dots correspond to places reached in July 24th. The dark orange dots correspond to places reached in July 30th. We can assume that the area containing lighter orange dots (Romilly-sur-Seine, Nogent-sur-Seine, etc.) could be the place where the fear started to spread, which is confirmed by the text.

[Insert Figure 17 here: Zoom on the old province of Champagne]

Finally, we created another map in Figure 18 which exposes the multiple routes found by our algorithm according to proximity heuristics that follows. This map reflects the different orientations in which the fear travels from Paris and expands throughout the old Champagne province.

[Insert Figure 18 here: Multiple routes of the fear spread out of the old Champagne province using a different base map]

6 Conclusions and perspectives

In this article we proposed a graph matching approach based on qualitative spatial relations in order to resolve SNE in narrative texts describing routes. Our approach is made up of two steps. We first build a RDF graph of routes by parsing beforehand the tagged text in order to extract SNE, cardinal directions and motion verbs. In each route, waypoints are labeled by the SNE mentions and described by their spatial context: preceding, simultaneous and following waypoints and spatial relations with other waypoints. Afterwards, the SNE are resolved by matching the spatial context of each waypoint with the spatial context of its candidates taken from a KB. We thus used a graph matching approach based on graph edit distance and implemented the A* Beamsearch algorithm.

.

We tested our approach on two excerpts from the book "La grande peur de 1789" by Georges Lefebvre. They deal with the spread of great fear in the southwest of France for the first and in the Champagne region for the second. We compared our results with those of REDEN approach and tested several values for the different parameters used by the algorithm in order to improve its results in terms of precision and recall.

The results obtained are better than an existing named-entity linking approach, but further comparisons with other approaches like DBpedia Spotlight or Perdido should be performed for further testing of the approach.

The first step of our approach could benefit from natural language processing methods in order to improve the routes RDF graph construction. As a matter of fact, we have noticed that some spatial relations could not be found by the heuristics we had implemented for this task. Moreover the three cost functions can certainly be improved, especially the deletion and insertion functions which are simple constant values so far. For now, each sequence of routes is considered as a context in which identical mentions of SNE necessarily refer to the same place. But it would be interesting to test the hypothesis "one sense per discourse" on the whole text since each extract concerns one and only one region, and therefore if a toponym is mentioned several times, it may be supposed that it designates the same place.

The source code is available in open source on GitHub[18], as well as the two reusable ontologies. We also found a problem of exhaustiveness of the number of qualitative spatial relations in DBpedia.

In addition to the complementary tests already mentioned, the cartographic module needs to be improved according to the needs of historians. Indeed, as it currently stands, it does not always make it possible to properly understand the different routes that make up a sequence, partly because of the manner in which the waypoint bags are managed. Many technical improvements can also be made to improve the performances of our algorithm (database for shortest paths, further parallelization, etc.).

We also found a need for DBpedia to be enriched with more spatial relations by using more classical geographical databases like we did in our work. Such resource does not currently exist. Instead, historical gazetteers such as Pleiades[19] focuses on the Ancient times. Fortunately, ongoing efforts engage at producing geo-historical data[20] that will be well-adapted to our spatial and temporal constraints. These data can easily be

[18] https://github.com/cvbrandoe/REDEN/tree/geo-reden
[19] http://pleiades.stoa.org
[20] https://www.geohistoricaldata.org

transformed into LD using existing tools (Hamdi et al. 2014) and integrated into our approach.

Acknowledgements

The authors would like to thanks Stéphane Baciocchi, Historian in the Historical Research Center at the School for Advanced Studies in the Social Sciences (EHESS) in Paris, for his help in the constitution of the corpus and for the interesting discussions and expertise on La Grande Peur.

References

Agrawal, R. J., and Shanahan, J. G. 2010. ‘Location disambiguation in local searches using gradient boosted decision trees’. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 129-136). ACM.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. 2004. ‘Web-a-where: geotagging web content’. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 273-280). ACM.

Batista, D. S., Ferreira, J. D., Couto, F. M., and Silva, M. J. 2012. ‘Toponym disambiguation using ontology-based semantic similarity’. Computational Processing of the Portuguese Language, 179-185.

Bensalem, I., and Kholladi, M. K. (2010). Toponym disambiguation by arborescent relationships. Journal of Computer Science, 6(6), 653.

Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S. V. N., Smola, A. J., and Kriegel, H. P. (2005). Protein function prediction via graph kernels. Bioinformatics, 21(suppl 1), i47-i56.

Blank and Henrich. 2016. ‘A depth-first branch-and-bound algorithm for geocoding historic itinerary tables’. In Proceedings of the 10th Workshop on Geographic Information Retrieval (GIR '16). ACM, New York, NY, USA, , Article 3 , 10 pages. DOI: https://doi.org/10.1145/3003464.3003467

Brando, C., Frontini, F., & Ganascia, J. G. (2015). Disambiguation of named entities in cultural heritage texts using linked data sets. In East European Conference on Advances in Databases and Information Systems (pp. 505-514). Springer International Publishing.

Brando, C., Frontini F., Ganascia, J-G. 2016. "REDEN : Named Entity Linking in Digital Literary Editions Using Linked Data Sets". Complex Systems Informatics and Modeling Quarterly 0 (7) : 60-80. http ://dx.doi.org/10.7250/csimq.2016-7.04

Buscaldi, D. 2011. ‘Approaches to disambiguating toponyms’. SIGSPATIAL Special, 3(2), 16-19.

Buscaldi, D., and Rosso, P. 2008. ‘A conceptual density-based approach for the disambiguation of toponyms’. International Journal of Geographical Information Science, 22(3), 301-313.

Buscaldi, D., and Rosso, P. 2008. ‘Map-based vs. knowledge-based toponym disambiguation’. In Proceedings of the 2nd international workshop on Geographic information retrieval (pp. 19-22). ACM.

Cohn, AG., Li, S., Liu, W., Renz, J. 2014. ‘Reasoning about topological and cardinal direction relations between 2-dimensional spatial objects’. J. Artif. Int. Res. 51, 1, 493-532.

Frank, A., 1996. ‘Qualitative spatial reasoning: cardinal directions as an example’. International Journal Of Geographical Information Systems Vol. 10 , Iss. 3

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. 2013. ‘Improving efficiency and accuracy in multilingual entity extraction’. InProceedings of the 9th International Conference on Semantic Systems (pp. 121-124). ACM.

Derungs, C., and Purves, R. S. 2014. ‘From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus’. International Journal of Geographical Information Science, 28(6), 1272-1293.

Ferragina, P., and Scaiella, U. 2010. ‘Fast and accurate annotation of short texts with wikipedia pages’. arXiv preprint arXiv:1006.3498.

Frontini, F., Brando, C., Ganascia, J-G. 2016. "REDEN ONLINE : Disambiguation, Linking and Visualisation of References in TEI Digital Editions". In Digital Humanities 2016 : Conference Abstracts, 193-197. Krakow : Jagiellonian University & Pedagogical University. http ://dh2016.adho.org/abstracts/362.

Gärtner, T., Lloyd, J. W., and Flach, P. A. 2002. ‘Kernels for structured data’. In International Conference on Inductive Logic Programming (pp. 66-83). Springer Berlin Heidelberg.

Habib, M.B. and van Keulen, M. 2012. ‘Improving Toponym Disambiguation by Iteratively Enhancing Certainty of Extraction’. In: Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2012, 4-7 Oct 2012, Barcelona, Spain. pp. 399-410. SciTePress. ISBN 978-989-8565-29-7

Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J. 2013. ‘Evaluating Entity Linking with Wikipedia’ Artificial intelligence, vol. 194, pp. 130–150,. Available: http://dx.doi.org/10.1016 /j.artint.2012.04.005

Hakimov, S., Oto, S. A., and Dogdu, E. 2012. ‘Named entity recognition and disambiguation using linked data and graph-based centrality scoring’. InProceedings of the 4th international workshop on semantic web information management (p. 4). ACM.

Hamdi, F., Abadie, N., Bucher, B., Feliachi A. 2014. ‘GeomRDF: A Geodata Converter with a Fine-Grained Structured Representation of Geometry in the Web’, in the first International Workshop on Geospatial Linked Data (GeoLD 2014) - SEMANTiCS 2014, 1 september 2014, Leipzig, Germany

Han, X., and Sun, L. 2011. ‘A generative entity-mention model for linking entities with knowledge base’. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 945-954). Association for Computational Linguistics.

Intagorn, S., and Lerman, K. 2011. ‘Learning boundaries of vague places from noisy annotations’. In Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 425-428). ACM.

Ireson, N., and Ciravegna, F. 2010. ‘Toponym resolution in social media’. The Semantic Web–ISWC 2010, 370-385.

Lefebvre, G., 1932. ‘La grande peur de 1789’, Paris, Armand Colin

Leidner, J. L. 2007. ‘Toponym resolution in text: annotation, evaluation and applications of spatial grounding’. In ACM SIGIR Forum (Vol. 41, No. 2, pp. 124-126). ACM.

Livi, L., and Rizzi, A. 2013. ‘The graph matching problem’. Pattern Analysis and Applications, 16(3), 253-283.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., & Gaio, M. (2014, November). Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 183-192). ACM.

Moncla, L. 2015. ‘Automatic reconstruction of itineraries from descriptive texts’ (Doctoral dissertation, Université de Pau et des Pays de l'Adour; Universidad de Zaragoza).

Moretti, F., 2007. ‘Graphs, Maps, Trees: Abstract Models for Literary History’. London, New York: Verso.

Moro, A., Raganato, A., and Navigli, R. 2014. ‘Entity linking meets word sense disambiguation: a unified approach’. Transactions of the Association for Computational Linguistics, 2, 231-244.

Neuhaus, M., Riesen, K., and Bunke, H. 2006. ‘Fast suboptimal algorithms for the computation of graph edit distance’. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR) (pp. 163-172). Springer Berlin Heidelberg.

Overell, S., and Rüger, S. 2008. ‘Using co-occurrence models for placename disambiguation’. International Journal of Geographical Information Science,22(3), 265-287.

Porter, Atkinson, Gregory. (2015) “Geographical Text Analysis: A new approach to understanding nineteenth-century mortality” Health and Place, 36, pp.25-34. See:doi:10.1016/j.healthplace.2015.08.010

Porter, Gregory, Atkinson. 2016 ‘Investigating the temporal and spatial representations of disease in nineteenth-century British newspapers through text analysis and GIS’, American Association of Geographers, San Francisco, USA, 30/03/2016.

Riesen, K., and Bunke, H. 2009. ‘Approximate graph edit distance computation by means of bipartite graph matching’. Image and Vision computing, 27(7), 950-959.

Riesen, K., and Bunke, H. 2010. ‘Graph classification and clustering based on vector space embedding’. World Scientific Publishing Co., Inc.

Santos, J., Anastácio, I., and Martins, B. 2015. ‘Using machine learning methods for disambiguating place references in textual documents’. GeoJournal, 80(3), 375-392.

Smith, D. A., and Crane, G. 2001. ‘Disambiguating geographic names in a historical digital library’. Research and Advanced Technology for Digital Libraries, 127-136.

Smith, D. A., and Mann, G. S. 2003. ‘Bootstrapping toponym classifiers’. In Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1 (pp. 45-49). Association for Computational Linguistics.

Speriosu, Baldridge, J. 2013. ‘Text-Driven Toponym Resolution using Indirect Supervision’. ACL (1), 1466-1476.

Usbeck, R., Ngomo, A. C. N., Röder, M., Gerber, D., Coelho, S. A., Auer, S., and Both, A. 2014. ‘AGDISTIS-graph-based disambiguation of named entities using linked data’. In The Semantic Web–ISWC 2014 (pp. 457-471). Springer International Publishing.

Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. 2011. ‘Aida: An online tool for accurate disambiguation of named entities in text and tables’. Proceedings of the VLDB Endowment, 4(12), 1450-1453.

Zhao, J., Jin, P., Zhang, Q., and Wen, R. 2014. ‘Exploiting location information for web search’. Computers in Human Behavior, 30, 378-388.